DOCUMENT RESUME

ED 400 331 TM 025 736

AUTHOR	Bridgeman, Brent; And Others
TITLE	Reliability of Advanced Placement Examinations.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-96-3
PUB DATE	Feb 96
NOTE	31p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Advanced Placement; *College Entrance Examinations; *Error of Measurement; Higher Education; High Schools; *High School Students; Interrater Reliability; *Scores; Statistical Analysis; *Test Reliability
IDENTIFIERS	*Advanced Placement Examinations (CEEB); Free Response Test Items; *Inconsistency

ABSTRACT

The various methods for computing the reliability of scores on Advanced Placement (AP) examinations are summarized. For the free response portion of the examinations, raters can contribute to score unreliability through both systematic severity errors (in which some raters consistently rate more severely than other raters) and through inconsistency. Inconsistency appears to be a much greater problem than systematic severity errors. Question-to-question variation (or score reliability) is seen as a greater problem than rater inconsistencies. The impact of increasing or decreasing the number of topics is demonstrated by showing the proportion of students correctly classified as the number of topics changes using the results of AP examinations for the 1993 school year. Procedures to enhance both rate and score reliability are discussed. A table of score reliability and correct grade classifications is presented as an appendix. (Contains 1 table, 3 figures, 1 appendix table, and 15 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 400 331

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

RESEARCH REPORT

RELIABILITY OF ADVANCED PLACEMENT EXAMINATIONS

Brent Bridgeman
Rick Morgan
Ming-mei Wang

Educational Testing Service
Princeton, New Jersey
February 1996

BEST COPY AVAILABLE

TM025736

RELIABILITY OF ADVANCED PLACEMENT EXAMINATIONS

Brent Bridgeman, Rick Morgan, and Ming-mei Wang

January, 1996

Copyright © 1996. Educational Testing Service. All rights reserved.

Abstract

The various methods for computing the reliability of scores on Advanced Placement examinations are summarized. For the free response portion of the examinations, raters can contribute to score unreliability through both systematic severity errors (in which some raters consistently rate more severely than other raters) and through inconsistency. Inconsistency appears to be a much greater problem then systematic severity errors. Question to question variation (or score reliability) is seen as a greater problem than rater inconsistencies. The impact of increasing or decreasing the number of topics is demonstrated by showing the proportion of students correctly classified as the number of topics changes. Procedures to enhance both rater and score reliability are discussed.

Key words: generalizability, reliability, advanced placement, essays

RELIABILITY OF ADVANCED PLACEMENT EXAMINATIONS

The Advanced Placement (AP) Program is a system of course descriptions and examinations that permits high school students to receive college credit and/or placement into advanced college courses for college-level courses taken in high school. The program is growing rapidly in both candidate volume and national recognition. In 1994, more than 458,000 students took one or more of the 29 examinations in the Advanced Placement (AP) program. These nationally standardized examinations are developed by committees of secondary school and college teachers in each subject area, and are administered in May of each year. They typically require about three hours of testing time. In addition to the traditional multiple-choice format, AP tests include essay (or complex problem-solving) questions that require extended constructed responses.

As with any test, it is important to demonstrate that the scores are reliable and to adopt procedures that will enhance the reliability of the scores. In the current context, reliability refers to the consistency of test scores. Once test specifications have been set, an examinee should receive approximately the same score regardless of which test form (particular set of questions) the examinee has taken or which scorers have rated the free-response questions. That is, scores should be as generalizable as possible across different test forms and different raters. Other sources of inconsistencies in scores, such as day to day variations caused by feeling better some days than others, could also be considered. Although such variation can be important, it is typically not under the control of the test developer. Because test developers can devise procedures that minimize form to form and rater to rater variability, we will focus primarily on these sources of unreliability.

AP examinations illustrate a difficult value tradeoff in assessment. The incorporation of different assessment formats and options enhances validity and fairness by creating an examination that is more representative of the course syllabus and the learning outcomes for individual students. On the other hand, it is well known that the use of optional material, performance samples, and free-response exercises will often yield lower reliability than would multiple-choice testing in the same amount of testing time. It can be argued that the unreliability introduced by including these free-response sections can work against the validity and fairness of credit or placement decisions based on AP grades. However, measuring the right mix of skills imprecisely may be better than a high precision assessment of only a limited portion of the domain of interest. Indeed, Messick (1989) identifies construct underrepresentation as a major threat to construct validity. Furthermore, if educational consequences of a particular test design are considered as an integral part of the validity of that examination, and if free-response examinations encourage desirable educational practices (such as giving students extensive instruction and practice with essay questions), then the free-response sections could be seen as necessary for a valid examination.

Such procedures as allowing examinees to choose from among several essay topics, thus allowing students to demonstrate their skills in content areas they know best, have the potential for increasing fairness. But because topics cannot be exactly equated in difficulty, some students may be disadvantaged because they selected a topic with particularly rigorous grading standards. With increasing interest and emphasis on diversity in assessment, it is important to develop assessment models and procedures that enhance both reliability and representativeness insofar as

6A

possible. Identifying and modeling sources of unreliability is a critical first step, but the ultimate goal is to produce an examination system that is more valid and fairer.

In the sections that follow, first we describe the procedures that are currently used for estimating the reliability of AP scores, and show the similarities and differences among the AP examinations on these indexes. Next, we describe methods for estimating and improving rater reliability. Finally, we describe methods for estimating and improving the reliability of the scores on the free-response tasks.

Current Procedures for Estimating Reliability on AP Examinations

For each of the multiple-component AP examinations, a composite score is derived as a weighted sum of the component scores (with weights reflecting the relative importance of each component in the subject area). The AP grades are reported on a five-point scale by dividing the composite scores into five intervals based on program-determined cut scores. The cut points are established by the Chief Reader using equated multiple-choice scores as a reference point. This score reporting system requires three steps to evaluate its reliability.

First, determine reliability of component scores. The reliability of each of the component scores must be determined directly with empirical data or indirectly by means of a measurement model. Typically, Kuder-Richardson formula 20 (with the Dressel adaption for formula scores) is employed as a lower-bound estimate of the reliability of a multiple-choice section.

Estimation of the reliability for the constructed-response component is not as straightforward. Because the free-response section consists of several prompts each scored by a different rater, both differences among raters and differences among topics contribute to error variance in scores. Coefficient alpha is employed as the lower-bound reliability estimate that

simultaneously includes both topic differences and rater differences as inseparable sources of error. This practice assumes that the topics are tau-equivalent (parallel) which may not be empirically or conceptually supported. Essay topics may be intentionally not parallel (e.g., the document based question and the standard question on the United States History examination). Therefore, the currently used reliability estimation methods for these intentionally non-parallel questions might be too low. We conducted a special study to explore this question in depth for the history examinations (see Bridgeman, Morgan, & Wang, 1996b). These examinations were targeted because they appeared to be especially prone to reliability underestimation by the usual methods; they exhibited a low correlation between the two essays which were explicitly created to tap different skills. However, the findings suggested that the current procedure does not underestimate the essay score reliability; the correlation of the scores on the essays written in response to the document based question and the standard essay question were just about as high as the correlations based on two document based questions or two standard essays. Although the history examinations were selected because they were thought to represent a worst case scenario, it is impossible to generalize these results to all of the other AP examinations.

Rater reliability cannot be routinely estimated separately from total free-response score reliability because each response is read by only a single reader. However, special reader reliability studies, in which each essay is read by at least two readers, are conducted periodically. Older studies used the correlation of first and second readings to establish reader reliability. Current practice uses a variance components analysis. Both methods yield very similar values. A standard error of measurement (due to raters) is computed for each question and these standard errors are multiplied by the question weights, squared and summed. This rater variance is then

A

divided by the total free-response score variance to find the proportion of variance due to reader inconsistency. The rater reliability is 1 minus this proportion.

Second, determine reliability of composite. An overall reliability estimate for the AP composite score is obtained by combining the component reliabilities, under the assumption that measurement errors are independent of each other and uncorrelated with the true component scores. Specifically, the error variance for each part score is multiplied by its weight (squared), the sum of these weighted variances is divided by the total composite score variance, and this product is subtracted from 1. This step seems to cause little technical concern except when examinee choice is implemented in the exam (i.e., when the examinee is free to choose from among two or more topics). The set of questions for choice may not be of equal difficulty, and when different questions are read by different raters, scoring stringency may vary across the choice questions. Thus, scores for the optional questions may not be comparable for examinees choosing different questions (Wainer, Wang, & Thissen, 1991; Pomplun, Morgan, & Nellikunnel, 1991). As examinee choice may be correlated with ability (and therefore true scores of other components), the reliability of the composite scores can be inappropriately estimated if the potential incomparability of the scores among choice questions is ignored. Given the thorny psychometric problems raised by choice, it should probably be discouraged in most circumstances. However, in some situations, choice may still be desirable. For example, if the examination is assessing the ability of students to organize evidence and present a cogent argument on a topic that they have studied in depth, choice may be required if different students have emphasized different topics in their studies. (See Bridgeman, Morgan, & Wang [1996a] for a more thorough discussion of this problem, including some data on the benefits of allowing choice.)

Third, determine the reliability of AP grades. The reliability of the composite scores is employed to estimate the decision accuracy and consistency of the grades (reported on a 1 to 5 scale). An algorithm originally devised by Livingston and Lewis (1991) and subsequently revised, RELCLAS-COMP Version 4, is applied to examine the reliability of classification of AP composite scores into the five-point grade scale and to describe the accuracy of the reported grade. This procedure permits estimation of the proportion of AP candidates whose estimated true grade is the same as their reported grade. The true grade is the grade that the candidates would receive if they could take a large number of alternative forms of the examination (with no learning from form to form) and their scores on all of these parallel examinations were averaged[1]. Perhaps more important than estimating these exact agreements, RELCLAS also estimates the proportion of examinees who are correctly classified as above or below each of the four possible cut scores (e.g., one cut compares student with a grade of 5 or 4 with students who score 3 or below). Many colleges grant credit for grades of 3 or higher. If a student whose true score is a 5 receives a 4 on the particular examination taken, this error is of little practical consequence to the student. But if a student with a true score of 3 mistakenly receives a 2, the consequences are significant. Thus, an examination in which exact agreement between reported grade and true grade is relatively low may still be relatively accurate in determining whether a student is above or below the 3/2 grade boundary. For example, on the 1993 English Literature and Composition examination, the proportion of examinees with exact agreement between true grade and reported

[1]Sometimes true score estimates include an allowance for day to day fluctuations in how an examinee feels, but the true score estimated here accounts only for form to form fluctuations.

grade was only .64, but the proportion correctly classified as above or below the 3/2 cut point was .89.

Summary of AP Reliability Information for the 1993 Test Year

A summary of the reliability information for the 1993 test year is presented in Table 1[2]. Examinations are listed in order by composite score reliability. In general, foreign languages and physical sciences dominate the top of the list (most reliable) while humanities and social sciences predominate at the bottom of the list. The ordering of the free-response reliabilities matches the ordering of the composite score reliabilities fairly closely. It has been asserted that the examinations with the most reliable free-response sections rely on analytical rather than holistic scoring (Wainer & Thissen, 1994), but this distinction is far from clear cut. The foreign language examinations use both analytic and holistic scoring, but the holistic scores contribute substantially more to the total for the free-response sections. For example, analytically-scored questions on the German Language examination account for only 21.5 out of 100 points on the free-response section, with the remaining 87.5 points coming from holistically-scored questions. In addition, the classification of a particular scoring rubric as holistic or analytic is somewhat arbitrary. Although the scoring of the essays in U.S. History and European History is usually classified as holistic, the scoring standards have several characteristics of analytical scoring schemes including quite specific standards on particular pieces of information or relationships that must be noted to obtain a high score.

[2]Studio Art-General and Studio Art-Drawing are omitted because they are not evaluated with traditional examinations. Instead, students submit structured portfolios that are organized into three sections. One section is scored by three different judges and the other two sections are each scored by two different judges. Composite score reliability was .89 for General and .91 for Drawing. Reliability issues related to the Studio Art portfolio assessment are discussed in considerable detail by Myford and Mislevy (1994).

BEST COPY AVAILABLE

The large number of free-response questions in the foreign language examinations reflects a number of very short tasks, not more time allocated for the free-response questions. For example, the German Language free-response section takes 70 minutes and includes 20 paragraph completion fill-in questions (with a suggested time limit of 10 minutes)[3], one composition (with a suggested time limit of 40 minutes), and seven questions in the speaking portion of the examination (20 minutes). In contrast, the U. S. History Examination allows 105 minutes for the free-response section (including a 15-minute reading period), but the examinee writes only two extended essays in this time[4].

The second column of the table indicates the percentage contribution of the free-response section to the composite score. This percentage is based on the total number of points contributed by each question type after score weighting. For example, in Biology, the 120 multiple-choice questions are weighted by .75 (for a maximum of 90 points) and the sum of the four essays (each scored on a ten point scale) is weighted by 1.5 (for a maximum of 60 points). Thus, the essay contributes 40% to the maximum possible score of 150. Although this nominal contribution could misrepresent the actual contribution to relative rankings if the variability of either score were severely curtailed[5], standard deviations of both sections are monitored to at least avoid gross distortions. Nominal contributions for the free-response score range from 33% (in

---

[3]In general, free-response sections have an overall time limit that is enforced by the test administrator. However, the time to be allocated to each activity within the section is merely suggested.

[4]Descriptions of test content and timing are for the 1993 examinations only. As a constantly evolving program, some of these details change from year to year. For example, in 1994, a third essay was added to the U. S. History examination.

[5]In the extreme case, suppose everyone got exactly the same score on the free-response section. Relative rankings would then be totally determined by scores on the multiple-choice test regardless of the nominal weight given to the free-response scores.

Psychology, Microeconomics, and Macroeconomics) to 60% (in History of Art and French Literature).

The next three columns of Table 1 summarize the score reliability estimates for the composite score, and separately for the multiple-choice and free-response sections. These score reliabilities include errors from both question to question and rater to rater variations; rater reliability by itself is estimated in the next column of the table.

For the examinations at the top of the list, the composite score reliability is typically higher than the reliability of the multiple-choice section alone. But for the examinations at the bottom of the list, reliability is higher for the multiple-choice section than for the composite. If maximizing reliability were the goal, the free-response scores for these examinations should be discarded. Of course, the goal is maximizing construct validity, not reliability. Because these essays measure something important that is not assessed by multiple-choice questions, even if they do not measure it very consistently, they make an important contribution to the assessment. Despite their lower reliabilities, the essay scores on the history examinations and the English Language and Composition examination predict college grades in those subjects as well as the more reliable multiple-choice scores (Bridgeman & Lewis, 1994).

As indicated in the rater reliability column, rater reliabilities are typically high compared to the overall reliability of the free-response scores (which includes both question to question and rater to rater variability) . Although there is room for improvement in the rater reliabilities (especially on the examinations at the bottom of the list), differences among tasks appear to make a greater contribution to score unreliability than do differences among raters. The critical

importance of variation across tasks and the relatively minor role of variation across raters has been consistently found in a number of different contexts (Shavelson, Baxter, & Gao, 1993).

The next to the last column in Table 1 indicates the percent of examinees whose reported grade is the same as their true grade (i.e., average grade if many different forms were taken). This percentage is quite modest for many of the examinations. On the English Literature and Composition examination, for example, only 64% of the candidates receive their true grade from the specific form that they happened to take. However, in terms of pass/fail decisions (if a grade of 3 or greater is considered passing), the reliability appears to be much more reasonable; 89% of the pass/fail decisions on a given form accurately reflect pass/fail decisions based on true grades. Given the broad range of exact agreements across the examinations (from 63% to 83%), the range of pass/fail agreements is relatively narrow (from 87% to 93%).

Methods to estimate and improve rater reliability

A number of separate components contribute to the unreliability of scores assigned by raters. One is reader severity; some readers consistently tend to give lower scores than other readers. The other is reader inconsistency; one reader may think the essay written by examinee A is better than the essay written by B and a different reader will think the essay written by B is better than the essay written by A. Statistical adjustments (or score calibration) are possible for reader severity, but there is no way to adjust for inconsistency. Braun (1988) reviewed the literature on the problem of scoring reliability and demonstrated a relatively simple statistical procedure for calibrating essay readers that substantially reduced errors due to reader severity. However, reader severity was a relatively unimportant contributor to score unreliability. Longford (1993) developed an additive variance components model that permits direct estimation

of reader characteristics and examinee's true scores. For both AP examinations that he studied (Biology and Studio Art), reader inconsistency was a much larger contributor to error in scores than was reader severity. For one of the four Biology essays studied, the severity variance was essentially zero, and for the other three essays, the inconsistency variance was from to two to eight times as large as the severity variance. In a subsequent study, Longford (1994) studied the Psychology examination, the English Language and Composition examination, and two Computer Science examinations. Once again, severity variance was small relative to inconsistency variance, although the relative size of these components varied considerably from one topic to the next. For example, on one Psychology essay, the variance components were 4.55 (due to true differences over examinees), 0.34 (due to rater severity), and 1.32 (due to rater inconsistency); on the other essay, they were 5.17, 0.02, and 1.09 respectively. For the eight free response questions from the two Computer Science examinations, estimated true variance ranged from 4.75 to 8.89 while the severity component was 0.01 or less for six of the eight questions (and 0.19 and 0.03 for the other two). In comparison, the inconsistency variance ranged from 0.41 to 0.70. Severity variance was relatively more important on the two English Language and Composition essays studied, but inconsistency variance was still over three times as large as severity variance.

With a student writing several essays each read by a different reader, severity errors would tend to be inconsequential for most people. Thus, adjusting for reader severity would have only minimal impact on overall score reliability. Myford and Mislevy (1994), in their study of the AP Studio Art program (in which seven readers contribute to a score for each portfolio), noted that "adjusting for reader effects would not materially improve the accuracy of scores for this program" (p. 45). Adjustment for reader severity improved reliability by only .006 (from .887 to

.893). They also attempted to identify background variables (such as years of teaching experience) that might predict reader severity, but found that the variables studied had a negligible impact on predictions of reader severity.

Even though severity errors make a relatively small overall contribution to score unreliability, two arguments can be made for adjusting scores. First, even though the impact of score adjustment is small on average, a few individuals can be significantly affected if they are unlucky enough to be rated by an especially severe judge on each (or most) of the essays that they write. For these unlucky individuals, small severity errors can accumulate rather than cancel each other out. Second, the adjustment process can be completed relatively quickly and inexpensively by computer. Compared to the cost of hiring additional readers, the adjustment is very cost-effective. However, on the negative side, the possible psychological impact of the adjustment process on the readers should be considered. If raters know that their scores will be adjusted, they may become more lax in applying the scoring guidelines. Failure to strictly follow the guidelines could lead to the more serious inconsistency errors for which no adjustment exists. A simple experiment could determine whether such psychological effects on readers exist to any practically significant extent.

Another argument against adjustment is that it could disadvantage individuals who write very strong essays that would receive the highest scores even from the strictest readers. If these essays are read by lenient readers, the adjustment process (which automatically deducts points from essays read by lenient readers) will unfairly assign them lower scores than they deserve. This problem could be solved by having the system flag scores that moved from one grade to another as the result of a score adjustment, and the essays identified could be reread at a special reading.

However, even with an efficient storage system, finding a few essays out of the tens of thousands administered for this special rereading could be a logistic nightmare (or perhaps just a bad dream if essays were stored electronically). An additional consideration is that score adjustment may be very difficult to explain to students and their teachers.

No statistical procedure can adjust for reader inconsistencies, but that does not imply that substantial inconsistencies must be accepted as inevitable. Inconsistencies can be reduced through more comprehensive rater training and monitored practice. Training would also reduce severity errors. Even if severity estimates were not used to adjust scores, they could be useful in the training and monitoring of raters. Raters who were identified in the first day of the reading as consistently too severe or too lenient could be retrained.

In 1992, the leadership of the AP English readings began to emphasize the need to improve reader training and reader monitoring. Over time this has resulted in more training for both table leaders and individual readers, an increase in the number of check readings performed by table leaders, more timely and better information given to table leaders concerning the scores given by each of their readers, and an emphasis toward the use of a common scoring standard and a corresponding emphasis away from the use of individual scoring standards. A comparison of the AP English reader reliability studies conducted five years apart (Maneckshana, Stevens, & Damiano, 1990) and (Maneckshana and Morgan, 1995) shows that the reader reliability of the free response section was more than .10 higher and that the average correlation between free response questions was nearly .10 higher for the later study. These findings illustrate the benefit that can be associated with improved reader training and monitoring.

Methods to estimate and improve free-response score reliability

Probably the easiest way to enhance score reliability is to add more tasks. In order to show the effects of changing the number of free-response tasks on score reliability, we selected four high-volume AP examinations that represented diverse content areas and differing levels of composite score reliability (Biology, Computer Science A, Psychology, and English Language and Composition). The Computer Science A and Biology examinations each had four free-response tasks, English Language and Composition had three free-response questions (essays), and Psychology had two. The composite score reliability was estimated for shortened versions of each exam, assuming that the weight on the free-response section remained constant as the number of free-response questions was reduced (40% for Biology, 50% for Computer Science A, 33% for Psychology, and 55% for English Language and Composition). For example, in Biology there were four estimates when one question was removed; one estimate assumed only essays 1, 2, and 3 had been administered, a second estimate assumed only essays 2, 3, and 4 had been administered, etc. Thus, there were six estimates for the exam with 2 of 4 free-response questions included, and four estimates for the exam with only one free response question. The composite score reliability estimates for each number of free-response questions were then averaged to find the estimated composite score reliability for an exam with each number of free response questions. These estimates are plotted in Figure 1. (See Appendix Table A-1 for the numbers used to generate this figure.)

Although the curve for English Language and Composition was noticeably below the other curves (because of the relatively low reliability of the free-response scores combined with the relatively large weight on the free-response section), the shapes of the curves were quite

consistent across the four examinations. Increasing from one to two free-response questions apparently substantially improves composite score reliability with lesser gains as more essays are added. Note that even the relatively unreliable English Language score, based on its regular complement of three essays, was as reliable as the Biology examination would be if it contained only a single essay. No AP examination relies on a single question in its free-response section, and these curves suggest that this is a wise policy from a reliability point of view.

Figure 1 is useful for showing general trends, but differences in the number of people correctly classified as the number of questions increases are not immediately apparent. The percent of examinees correctly classified thus provides useful additional information. The procedure to identify classification errors used the composite score reliability, the standard error of measurement, and the distribution of observed scores for each shortened exam version as input to RELCLAS Version 4. Cut points for each shortened examination were estimated based on the equal percentile equivalents to the proportion of AP candidates below each cut point in the full exam. The RELCLAS program estimated the proportion of examinees whose AP grades were correctly categorized. These estimates were averaged to find the RELCLAS reliability estimates for the shortened forms. The percent of correct classifications for different numbers of free-response questions are presented in Figure 2.

The percent of examinees whose grades were correctly categorized simply in relation to the 2/3 cut point is presented in Figure 3. As noted previously, the percent of examinees correctly classified as either 3-or-above or as 2-or-below is substantially above the percent of examinees who receive the exactly correct grade. Figure 3 also suggests that if correct classification at this cut point is the primary goal, little is to be gained by including more than two or three free-

response exercises. For example, in Biology, going from one essay to two results in about 3 more correct classifications for each 100 students, but going from three to four essays results in only one additional correct classification. On the other hand, this one additional correct classification (per 100 students) represents a full 10% of the misclassified students. Note that these calculations are all based on the assumption that about half of the score (or more) is derived from a highly reliable multiple-choice section. If these multiple-choice questions were not included, the effect of adding a third or even a fourth free-response exercise would be much more dramatic.

Summary

The reliability of various components of AP examinations was estimated through a variety of techniques including correlation-based and variance-components methods. Regardless of method, a number generalizations are possible. The reliability of the multiple-choice section and the composite score for all examinations was high (at least .82). There was considerable variation in the reliabilities of the free-response sections (from a high of .88 in Computer Science A to a low of .49 in European History). Free-response scores in foreign languages, science, and mathematics tended to be more reliable than essay scores in the social sciences and humanities. Rater reliabilites were also quite variable, ranging from a high of .98 in Computer Science A to a low of .67 in U. S. History. Research reviewed suggested that rater inconsistency was a much greater problem than raters who provided ratings that were consistently too high or too low. Topic to topic variability was more important than rater to rater variability. Rater reliability can be improved by enhanced training, and free-response score reliability can be enhanced by adding topics. Although there was considerable variation in the estimated percent of students correctly

classified into the appropriate 1-5 AP grade (from 63% to 83%), the percent correctly classified as either 3-or-above or 2-or below was uniformly high (from 87% to 93%).

References

Braun, H.I. (1988). Understanding score reliability: Experience calibrating essay readers. Journal of Educational Statistics, 13, 1-18.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. Journal of Educational Measurement, 31, 37-50.

Bridgeman, B., Morgan, R., & Wang, M. (1996a). Choice among essay topics: Impact on performance and validity. Princeton, NJ: Educational Testing Service.

Bridgeman, B., Morgan, R., & Wang, M. (1996b). The reliability of document-based essay questions on Advanced Placement history examinations. Princeton, NJ: Educational Testing Service.

Livingston, S.A., & Lewis, C.L. (1991). Estimating the Consistency and Accuracy of Classifications Based on Composite Scores (version 4). Draft Report, Educational Testing Service.

Longford, N. T. (1993). Reliability of essay rating and score adjustment (ETS Technical Report 93-36). Princeton, NJ: Educational Testing Service.

Longford, N. T. (1994). A case for adjusting subjectively rated scores in the Advanced Placement tests (ETS RR- 94-58). Princeton, NJ: Educational Testing Service.

Maneckshana, B., Stevens, J., & Damiano, M. (1990). English Language and Composition Reader Reliability Study (SR 90-104). Princeton, NJ: Educational Testing Service.

Maneckshana, B., & Morgan, R. (1995). English Literature Reader Reliability Study (SR 95-08). Princeton NJ: Educational Testing Service.

Messick, S. (1989). Validity. In R. L. Linn (Ed.) Educational Measurement (3rd ed., pp. 13-104). New York: Macmillan.

Myford, C., & Mislevy, R. (1994). Monitoring and improving a portfolio assessment system. Princeton, NJ: Educational Testing Service.

Pomplun, M., Morgan, R., & Nellikunnel, A. (1992). Choice in Advanced Placement tests (SR-92-51). Princeton, NJ: Educational Testing Service.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. Journal of Educational Measurement, 30, 215-232.

Wainer, H., & Thissen, D. (1994). On examinee choice in educational testing. Review of Educational Research, 64, 159-195.

Wainer, H., Wang, X., & Thissen, D. (1991). How can we equate test forms that are constructed by examinees (RR-91-57). Princeton, NJ: Educational Testing Service.

Table 1

Summary of Reliability Statistics for 27 AP Examinations

Examination	Number of Free-Response Questions	% Composite From Free-Response	Reliability				% Agreement Reported and True Grade	
			Composite	Multiple-Choice	Free-Response	Rater	Exact	3/2 Boundry
German Language	28	50	0.94	0.94	0.82	0.92	83	93
Music Theory	6	50	0.93	0.91	0.85	–	77	93
Physics B	6	50	0.93	0.91	0.83	–	74	92
Computer Science A	4	50	0.93	0.86	0.88	0.98	73	92
Biology	4	40	0.93	0.94	0.76	0.89	76	92
Calculus AB	6	50	0.92	0.88	0.84	–	72	93
French Language	28	50	0.92	0.92	0.78	–	75	92
Physics C: Elec. & Mag.	3	50	0.92	0.84	0.86	–	71	92
Computer Science AB	4	50	0.91	0.86	0.82	0.97	69	91
Chemistry	9	55	0.91	0.90	0.81	–	73	91
Calculus BC	6	50	0.90	0.86	0.77	–	69	92
Physics C: Mechanics	3	50	0.90	0.86	0.78	–	70	93
Microeconomics	3	33	0.90	0.90	0.62	–	69	91
Spanish Language	27	50	0.89	0.89	0.70	0.81	71	90
Latin: Vergil	4	55	0.89	0.85	0.78	–	67	90
Latin: Catullus & Horace	4	55	0.89	0.88	0.74	–	66	91
Macroeconomics	3	33	0.89	0.89	0.55	–	67	91
History of Art	9	60	0.88	0.90	0.73	0.82	69	91
Psychology	2	33	0.87	0.90	0.57	0.86	66	90

BEST COPY AVAILABLE

Table 1, continued

Examination	Number of Free-Response Questions	% Composite From Free-Response	Reliability				% Agreement Reported and True Grade	
			Composite	Multiple-Choice	Free-Response	Rater	Exact	3/2 Boundry
Spanish Literature	3	55	0.85	0.87	0.69	0.75	68	91
Comparative Gov. and Pol.	2	50	0.85	0.89	0.63	--	66	89
French Literature	2	60	0.85	0.88	0.68	--	63	90
U. S. Gov. and Politics	2	50	0.84	0.88	0.54	--	66	88
European History	2	50	0.84	0.91	0.49	0.71	68	89
English Lang. and Comp.	3	55	0.84	0.85	0.63	0.68	67	87
U. S. History	2	50	0.83	0.90	0.53	0.67	66	87
English Lit. and Comp.	3	55	0.82	0.83	0.66	--	64	89

BEST COPY AVAILABLE

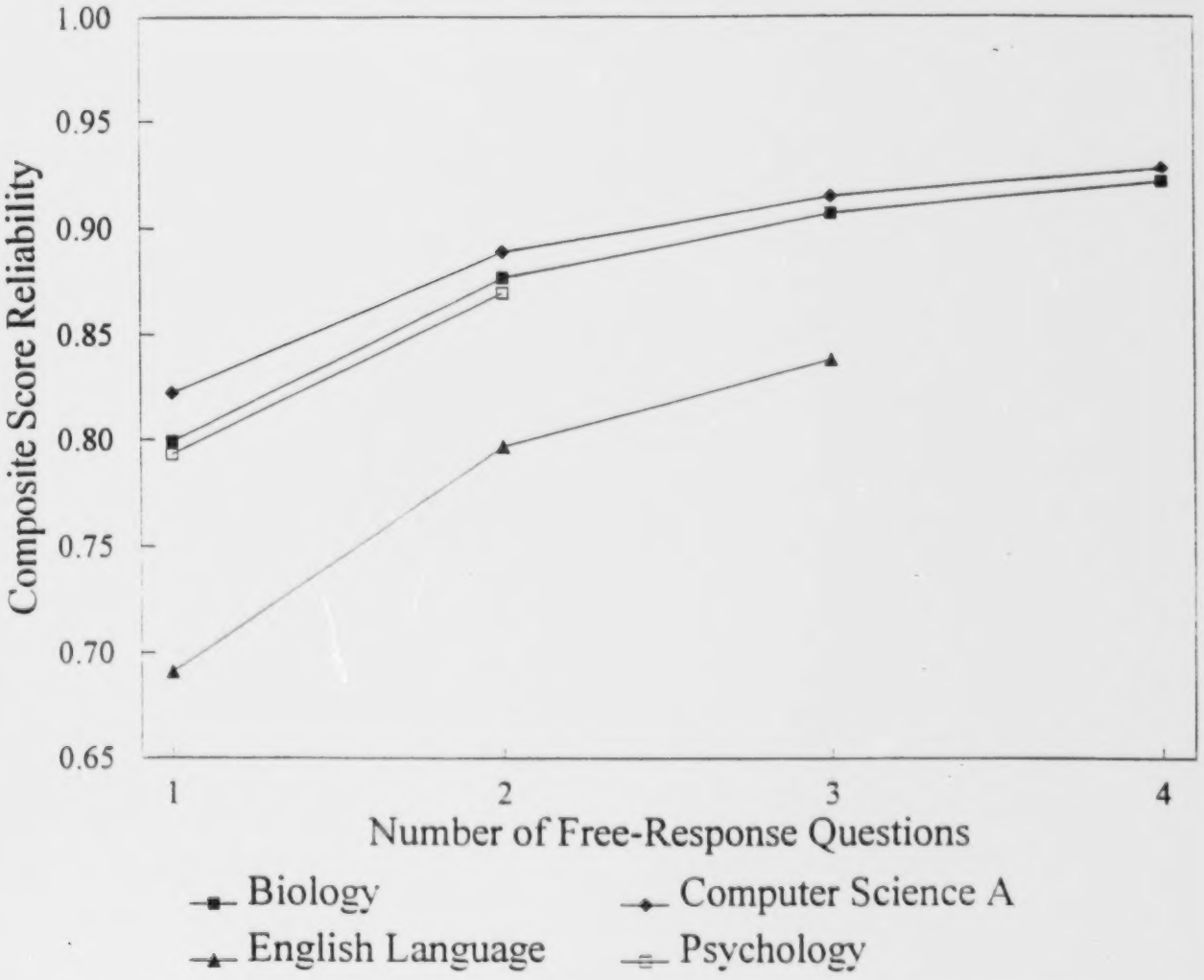

FIGURE 1. *Composite Score Reliability for different numbers of free-response questions*

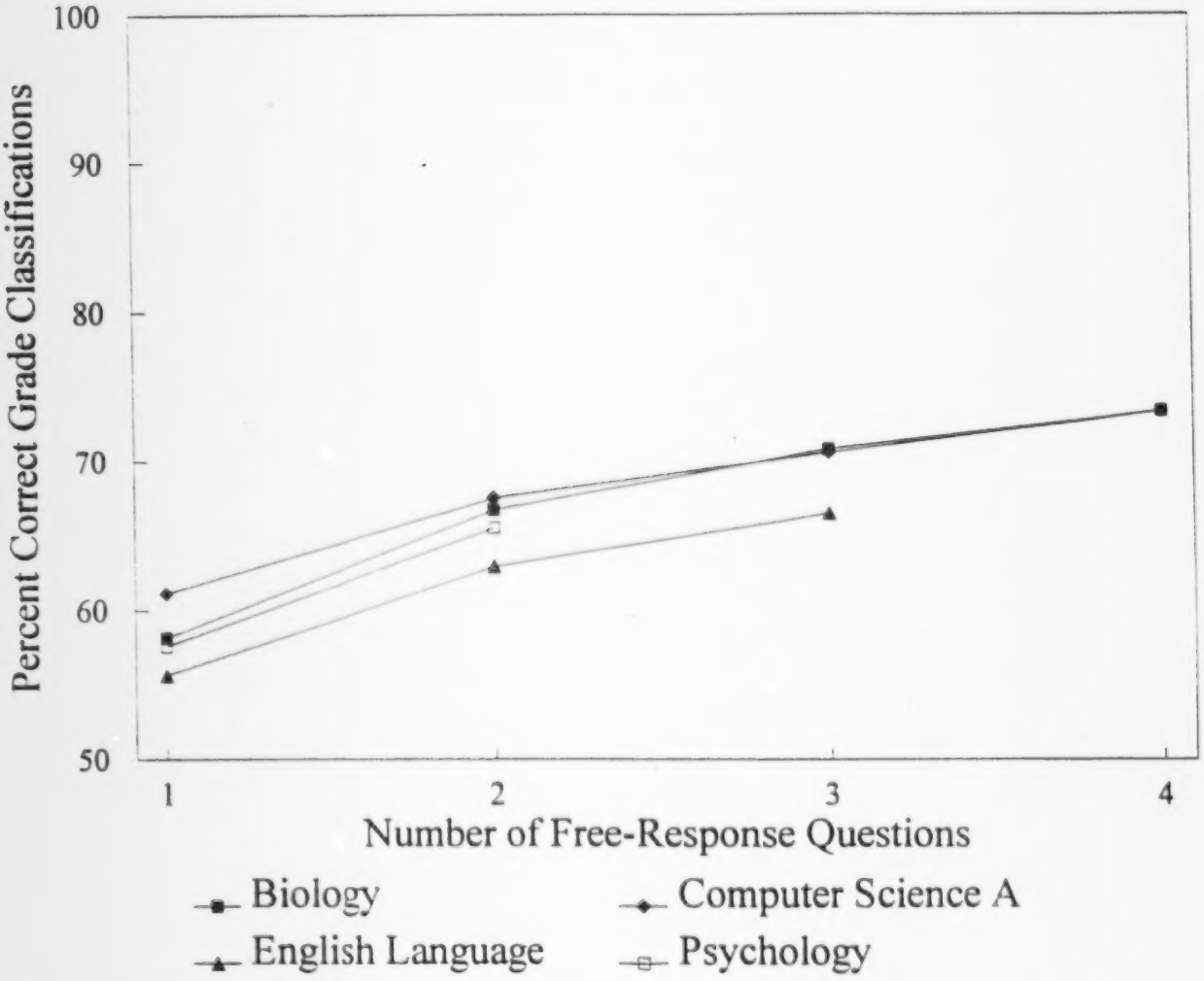

FIGURE 2. *Percent correct grade classifications for different numbers of free-response questions*

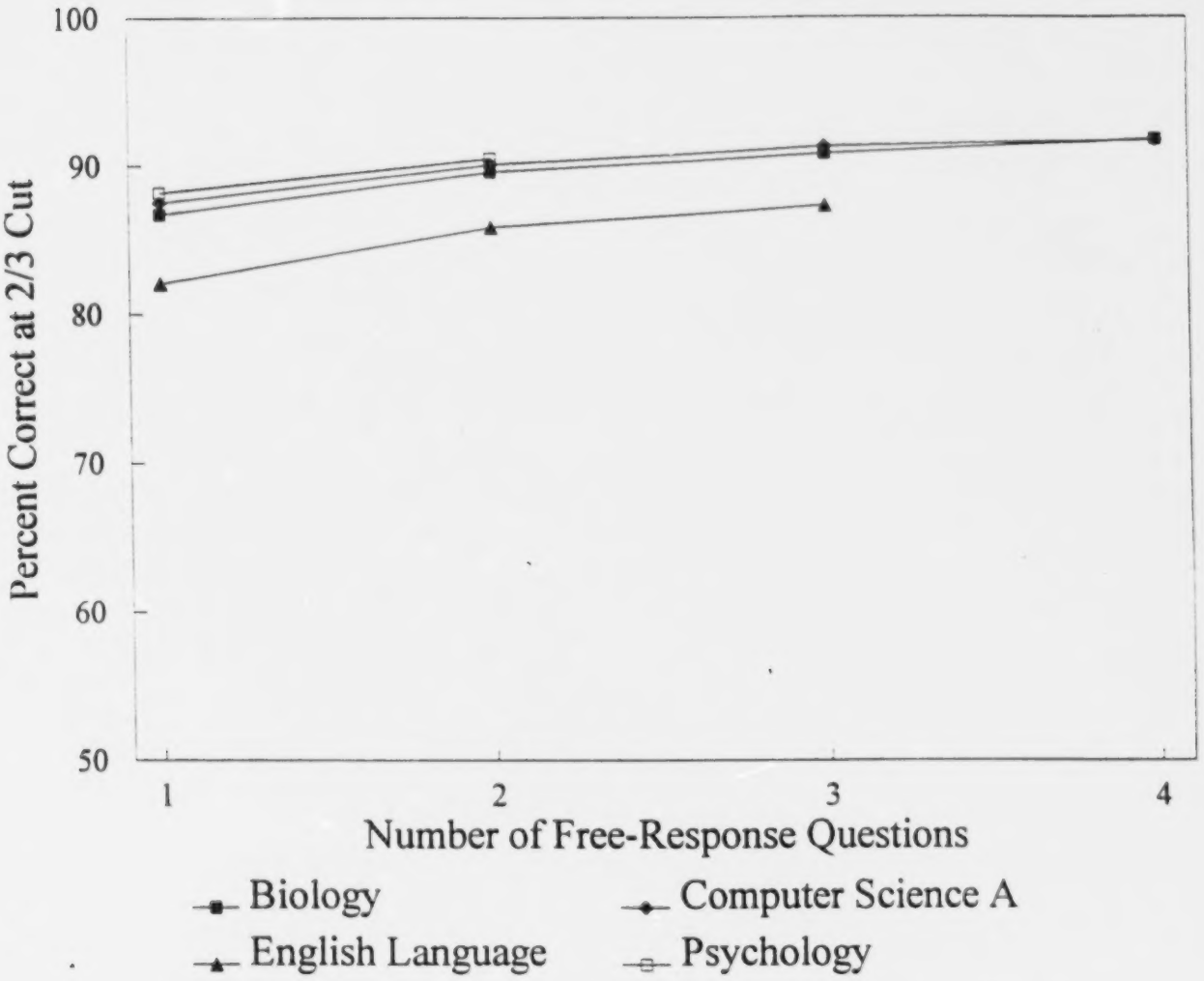

FIGURE 3. *Percent correct at 2/3 cut for different numbers of free-response questions*

Table A-1

Estimated Composite Score Reliability and AP Grade Correct Classifications as a Function of the Number of Free-Response Questions

Number of Questions	Composite Score Reliability	Percent Correct Grade Classifications	Percent Correct at 2/3 Cut
		Psychology	
1	.793	57.6	88.2
2	.870	65.5	90.5
		English Language and Composition	
1	.691	55.7	82.1
2	.797	62.9	85.9
3	.838	66.5	87.3
		Computer Science A	
1	.822	61.1	87.5
2	.889	67.5	90.1
3	.915	70.6	91.3
4	.928	73.4	91.7
		Biology	
1	.799	58.2	86.7
2	.877	66.7	89.6
3	.907	70.8	90.8
4	.922	73.4	91.8

Table A-1

Estimated Composite Score Reliability and AP Grade Correct Classifications as a Function of the Number of Free-Response Questions

Number of Questions	Composite Score Reliability	Percent Correct Grade Classifications	Percent Correct at 2/3 Cut
		Psychology	
1	.793	57.6	88.2
2	.870	65.5	90.5
		English Language and Composition	
1	.691	55.7	82.1
2	.797	62.9	85.9
3	.838	66.5	87.3
		Computer Science A	
1	.822	61.1	87.5
2	.889	67.5	90.1
3	.915	70.6	91.3
4	.928	73.4	91.7
		Biology	
1	.799	58.2	86.7
2	.877	66.7	89.6
3	.907	70.8	90.8
4	.922	73.4	91.8